Manuscripts in Time and Space: Experiments in Scriptometrics on an Old French Corpus*

Jean-Baptiste Camps

Centre Jean-Mabillon
École nationale des chartes | Paris Sciences & Lettres
E-mail: jbcamps@hotmail.com

Abstract

Witnesses of medieval literary texts, preserved in manuscript, are layered objects, being almost exclusively copies of copies. This results in multiple and hard to distinguish linguistic strata – the author's *scripta* interacting with the *scriptae* of the various scribes – in a context where literary written language is already a dialectal hybrid. Moreover, no single linguistic phenomenon allows to distinguish between different *scriptae*, and only the combination of multiple characteristics is likely to be significant [9] – but which ones? The most common approach is to search for these features in a set of previously selected texts, that are supposed to be representative of a given *scripta*. This can induce a circularity, in which texts are used to select features that in turn characterise them as belonging to a linguistic area. To counter this issue, this paper offers an unsupervised and corpus-based approach, in which clustering methods are applied to an Old French corpus to identify main divisions and groups. Ultimately, scriptometric profiles are built for each of them.

1 Introduction

Study on the diatopic variation of medieval French texts rests on the distinction proposed by Remacle [20] between *scripta*, written language (*Schriftsprache*), and dialect, spoken language, the latter mostly inaccessible to us. Based on his study of Walloon, this distinction was put forward as a mean to reconcile the difference he observed between the very characterized modern dialect and the medieval written texts from the area, presumably less marked by local traits. In the medieval *scripta*, he argued, the distinctive traits inherited from spoken Walloon would be present only by mistake or ignorance. Consequently, he formulated the apparently self-contradictory hypothesis that "1. the scripta was the result of a local development, 2. the scripta was a common language whose essential elements were found

*A digital appendix to this paper is available on Zenodo, DOI: 10.5281/zenodo.1117924. My gratitude, for discussion on this subject over the years, goes to Frédéric Duval, Martin D. Gleßgen, Hans Goebl and Achim Stein. I also thank the anonymous reviewers for their insightful advice.

in most spoken dialects of the *langue d'oïl*" (my translation). This distinction is now commonly accepted though sometimes criticised because it sets in stone our inability to ever gain insights into the reality of medieval dialects [4]. For the scholar who wants to date and localise the *scripta* of medieval texts, this implies that he will face a language that was never spoken as such and the very building blocks of which might be made of elements taken from various dialectal areas, maybe even a *koinè*, in which truly local traits are only marginal [7, p. 40].

The exact reality of this notion of *scripta* is still debated, but, as a working definition, we will take it as the written language, practised by a restricted number of literates, around scriptural centres (e.g. chancelleries), and supposedly conceived to allow for a broader comprehension than oral dialects, but still containing traits that can be geographically assigned to a specific area. The possible connexion between the main modern dialectal areas (as delimited by modern dialectologists) and the geographical hold of medieval documentary *scriptae* can be estimated due to the fact that administrative documents (charters, for instance) are usually dated (time and place date). It seems confirmed by Goebl's work [12].

The case is even more complex for literary witnesses[1]. While documentary texts (charters, wills,...) are practical documents, often of only local interest, most literary texts were made to be able to circulate through different linguistic areas, written by the more knowledgeable amongst the population, and influenced by the written codes of Latin [7, p. 41]. Sociolinguistics played a part, as well as factors related to production of books, such as the implantation of workshops. Variation in prestige between dialects led to difference in behaviour among writers, up to the point where some *scriptae* were judged distinctive of a genre, and its features imitated, like Western dialects or Picard for epic texts [1]. Two scribes working in the same workshop but from different origin might produce a text with different features. As such, localising the *scripta* of a witness does not mean as much finding its place of origin as identifying the linguistic inclinations of its writers [26]. But the major difficulty is of another nature yet: literary witnesses are layered objects, in which the language of the author interacts with each scribe's, up to the point where it is a very delicate task to assign any trait to a given layer, especially since any layer might already have included an alternation of forms or mixed forms [20].

As a consequence, it is very hard for dialectologists to determine isoglosses, or more precisely isographs [17, p. 166], that could clearly separate different *scriptae*. In fact, it is likely that no single trait can be used to define a *scripta* [9, p. 315]: most isographs are shared among several – usually neighbouring – regions [14, p. 65]. Even for the rare isographs that would be very distinctive, the information they provide is blurred by the hybrid nature of *scriptae* or the stratification of textual

[1] I define *witness* as a given instance of a text, as preserved in a particular document (usually, a manuscript) that is accessible to us. See Duval [6] for an account on the meaning of the terms *text* and *witness* ("texte" and "témoin") in (neo-lachamannian) textual criticism. It allows me to distinguish between the more abstract work (e.g. the story of Roland and the battle at Roncevaux) and its expression in particular texts (i.e. the *Chanson de Roland* or the *Cân Rolant*), attested in witnesses (e.g. *O*), preserved in documents (the ms. Digby 23).

witnesses. As a consequence, only a combination of traits, individually common with other *scriptae*, in a given relative frequency, makes the distinction possible. This has led to an emphasis put on quantification, and eventually on statistical multivariate analysis [9, p. 317]. This approach is named "dialectometry" since Séguy [22], or, better in our case, "scriptometry". It is defined by Goebl [10, p. 60-61] as an alliance between linguistic geography and clustering, and it shares some similarities with, for instance, stylometry and other historical text analysis fields. More generally, it can be defined as *the measure of scriptologic features*. As an exploratory approach, its goal is to reveal underlying structures that escape close reading analysis and are supposed to be more important that the superficial structures visible in the traditional maps of linguistic atlases [10, 11].

The dialectometric work of Dees or Goebl have been mostly founded on the listing of lexical, phonological or morpho-syntactical traits ("taxation" [10]), and the analysis of the resulting data. The atlases produced by Dees' team [3, 5] so include a series of maps that each present a quantified opposition between two groups of forms, and can be used [11, 3] as a matrix for computational analysis, both to study the underlying structures of dialectal variation or to locate a new text by confrontation with the already localised ones or to cartography similarities between regions and map dialectal areas [12, 3, 4, 5].

The work of Dees and his Amsterdam School and, after him, of Goebl and the Salzburg School, have given the rise to a more systematic and objective way to study medieval *scriptae* (for an historical synthesis, see Volker [27, chap. 2, p. 9-79]). Yet, an issue of circularity might still exist, since previous analyses usually based themselves on the localisation assigned to witnesses to identify linguistic areas and scriptological features. I would like to suggest a less supervised approach to the scriptometric analysis of the witnesses of a specific Old French epic genre, the *chansons de geste*. My aim will be to identify main divisions in the corpus and to create profiles for each of them, and to verify both customary separations between *scriptae* and the belonging of each individual witness to one of them.

2 Corpus and Method

In order to limit biases caused by stylistic, thematic or generic variations, this study will be limited to a single genre, the *chansons de geste*. Previous exploratory analyses, not shown here, on a multi-generic corpus of 299 texts, did confirm that generic differences interacted with linguistic boundaries and created too much noise. Authorship related biases are hard to avoid, but might be counteracted by the very graphic variation observed in the witnesses, a problem in the stylometric analysis of medieval vernacular texts. The corpus of *chansons* used here is composed of 50 witnesses (see app. A), with 1 104 296 tokens (geometric mean, 12 016, median, 11 490; min., 387; max., 217 942). The tokens are distributed between 52 202 forms (long-tail distribution, with 25 811 hapaxes; geom. mean of 2,57 occurrences, median, 2; 3rd quartile, 4). Editions were chosen for their use of a base

witness ("copy-text") – the emphasis here being on the witnesses and not on the original text – as well as for their availability in digital form. The selection of witnesses was done empirically to have the largest corpus with a representativity of several putative regions of origin. Yet, its heterogeneity is a limitation[2].

Variation in editorial practice regarding the allographs **i/j** or **u/v** and their transcription led me to map all of them on **i** and **u**. More generally, to avoid interferences with paleographic variation and perform on the graphematic level, all allographs (including "capitals") were normalized and all abbreviations expanded. The latter might be problematic, as it makes the process dependent on the choices of the editors, and can induce a bias, given that the norm is to use the majority unabbreviated form for expansion, inducing a distorsion favorable to this majority form as compared to the coexisting alternative ones [18, p. 33].

It is to be noted that the exclusion of allographetic variation is an important simplification of the reality of textual witnesses, done both for contextual (the unavailability of consistent information) and theoretical reasons, based on the assumption that the variation in use of variant letter forms is more dependent on scribe's idiosyncrasies or script variation (*textualis*, *cursiva*, etc.), sometimes termed "scribal mode" [15, 16]. In the terminology offered by McIntosh for his "scribal profiles", this means we will restrict ourselves to the "linguistic" by opposition to the "graphetic" components [15], that is "graphematic", opposed to "allographetic" in the terminology retained here [25]. Yet, given the interest of this latter kind of variation for dating and localising witnesses or identifying scribes, I have undertaken elsewhere to build a corpus of allographetic transcriptions and analyse them using similar techniques[3]. Another dimension of these witnesses that we will not take into account concerns the alterations to the content of the text during its transmission (variants), that is the way in which the behaviour of the scribe alters the text of his model to result in a new copy, that we could term the "diasystemic" component, after Segre's definition [21].

If previous scriptometric works were based on the "taxation" of a defined list of features, I chose to use a bag-of-words approach on the graphic forms of the texts, in order to avoid inducing *a priori* the features of the profiles. The main drawback is that occurrences of an identical phenomenon (e.g. graphs of a given diphthong) will be divided between all the forms that attest it. It will also prevent any syntactic feature to be taken into account and will limit the analysis to graphic or morphologic features. On the other hand, more limited habits, on the particular graph of a given lemma, will be fully accounted for. Lexical variation, important for the localisation of texts through the identification of regional words [7, p. 93],

[2] I intend to work, in the coming years, on the constitution of a corpus as exhaustive as possible of epic witnesses (transcriptions, critical editions, manuscript descriptions). The first few texts, encoded in TEI XML, are available on Github [8]. The data, in `csv`, used for this paper, are available with scripts to reproduce analysis, on the Zenodo repository.

[3] More details can be found in [2, chap. 2], including unsupervised clustering and allographetic profiles (sect. 2.4), with a digital appendix giving access to the datasets and analysis procedures. An updated version of the corpus is available in [8].

will also be analysed this way, even if it makes the analysis highly dependent on content-based variation. For this last reason, the database will be constituted of word rather than n-grams frequencies.

To limit content-based biases (and issues related to the non-Gaussian distribution of word-frequencies), only the most frequent words (MFW) are retained for analysis, an approach common in stylometry as well. Proper names were removed. This selection also leads to focusing the analysis on the dominant linguistic stratum (scribal or otherwise). Since no precise guidelines exist on the number of MFW to retain, robustness of the results will be checked with different levels of selection.

To cluster the witnesses, hierarchical clustering was retained, a common analysis in scriptometrics [10, 12].We do not yet possess guidelines on the effectiveness of various linkage criteria or distance measures in this field. Experimenting with a variety of those, to retain the one that would seem the best to me, though a heuristic approach advocated by Goebl [10, p. 85], would induce a validation bias. As a consequence, I retained Ward's method, because it relies on the barycentre of the data clouds and allows for the constitution of balanced and coherent clusters, often referred to as *types*, as it minimises intra-cluster variation and maximises inter-cluster variation [24]. It is usually claimed that only squared euclidean distance is correct to use with Ward's linkage, because it relies on computations in euclidean space. Yet, recent research by Strauss and von Maltitz [24] seems to demonstrate that it can be generalised to use with Manhattan distance, and that this metrics outperforms euclidean in what regards the classification of (indo-european) languages, a statement that agrees with previous research in computational phonology applied to the clustering of (Dutch) dialects [19], or with the supposed greater efficiency of Manhattan distance with highly dimensional data.

3 Results

Results were mostly stable with between 600 and 3000 MFW, as well as the agglomerative coefficient (between 0.83 and 0.8). The main divisions (fig. 1) are consistent with scriptological knowledge[4]. The first opposes supposedly Anglo-Norman witnesses to Continental ones. Inside the Anglo-Norman group, a division opposes older (XII or XIII$^{1/2}$) to more recent (XIII-XIV) witnesses, arranged in an imperfect chronological order. The orientation is in itself interesting as it seems to confirm the hypothesis that later Anglo-Norman texts, written in a fossilising linguistic context, were more subject to continental norm. The diachronic division of the Anglo-Norman group might also reveal the weakness of diatopic variation in this *scripta*, in a country where "*Normannica lingua, que adventitia est, univoca*

[4]Following preliminary experiments, a few too short (<2000 words) witnesses were removed, because their inclusion tended to slightly twist the analysis. Nonetheless, their placement was consistent with the rest of the clustering: Asprem_C was placed in the Anglo-Norman cluster, among witnesses from the middle of the XIIIth century, at an intermediary position between witnesses of earlier or later texts, just on the left of MacaireAl2B, whose placement was also consistent with chronology; the CharroiSch_fragm was in the Southern Lorraine group, with CharroiSch_D and PriseCordD; Fier_V was in the Lorraine/Burgundy group. See the online appendix.

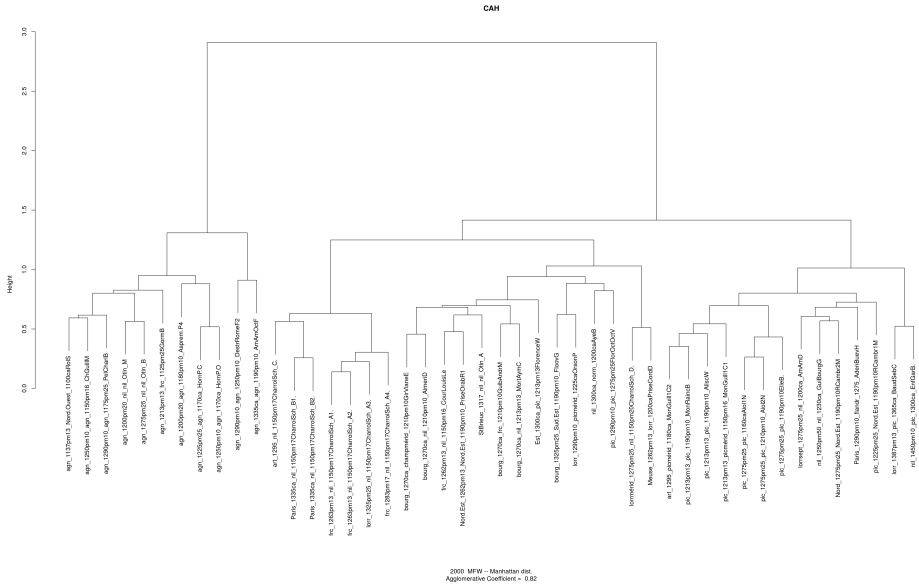


Figure 1: Hierarchical clustering of the *Geste* corpus (Ward's method, Manhattan dist., 2000 MFW, relative freq.)

maneat penes cunctos" (Ranulf Higden, *Polycronicon*, lib. I, cap. 59). The second division, considerably lower, creates a separation within the continental groups, namely dividing Picard witnesses of Picard texts from the rest.

The third division isolates mostly Central witnesses, but might also be due to authorial attraction between copies of the same text, that are even distributed between witnesses of the *A*, *B* and *C* versions (not the *D*). This might nonetheless have a linguistic sense, since *A1* and *A2* (and probably *A4*), for instance, are known to come from the same workshop [26, p. 434-436], as well as *B1* and *B2*.

Inside the group containing the rest of the Continental witnesses, which are mostly Eastern (or Lotharingian), divisions are weaker. Nonetheless, three subgroups can be individuated: witnesses from southern Lorraine (right), Burgundy (left), and Lorraine (centre). Many of the apparent exceptions can be explained and concern witnesses whose origin is subject to debate or need rectification. A subgroup of witnesses from Northern Lorraine or North-East appeared in the centre of this subgroup on some of the analyses (AmAmD, GuiBourG, RCambr), but are here grouped with Picard witnesses, maybe because one of them (RCambr) is a Northern copy of a text from the North-East.

Once groups are constituted, linguistic profiles for each of them can be built, at different levels, by estimating which features are the most characteristic with the values-test described by Lebart, Morineau et Piron [13, p. 181-184][5], giving us an

[5]The values-test is done by comparing $\bar{X}_k$, the mean of variable X in category k to the overall mean $\bar{X}$, while taking into account the variance $s_k(X)$ of this variable inside the class: $t_k(X) = \frac{\bar{X}_k - \bar{X}}{s_k(X)}$.

	v.test	mean in cat.	overall mean	sd in cat.	overall sd	p.value		v.test	mean in cat.	overall mean	sd in cat.	overall sd	p.value
Group 1 (*Anglo-Norman*)							Group 4 (*Picard*)						
pur	5.8438	0.0067	0.0018	0.0026	0.0032	0	ains	5.6322	0.0016	0.0005	0.0003	0.0007	0
sunt	5.7222	0.0058	0.0016	0.0024	0.0028	0	tous	5.4891	0.0021	0.0006	0.0006	0.0010	0
ad	5.6188	0.0120	0.0031	0.0056	0.0060	0	passes	5.2743	0.0002	0.0000	0.0001	0.0001	0
mei	5.5343	0.0019	0.0005	0.0010	0.0010	0	chou	5.2216	0.0009	0.0002	0.0006	0.0005	0
sur	5.5101	0.0044	0.0012	0.0021	0.0022	0	trestous	5.0875	0.0003	0.0001	0.0001	0.0002	0
lur	5.4663	0.0040	0.0010	0.0021	0.0021	0	tout	5.0120	0.0043	0.0015	0.0010	0.0020	0
tut	5.4522	0.0045	0.0012	0.0023	0.0023	0	sarrasins	4.9654	0.0004	0.0001	0.0003	0.0002	0
al	5.3361	0.0072	0.0022	0.0034	0.0036	0	sains	4.9536	0.0004	0.0001	0.0002	0.0002	0
e	5.3131	0.0357	0.0108	0.0127	0.0179	0	toutes	4.9496	0.0004	0.0001	0.0001	0.0002	0
sun	5.2683	0.0070	0.0018	0.0041	0.0037	0	commanda	4.9074	0.0001	0.0000	0.0001	0.0001	0
seit	5.2186	0.0020	0.0006	0.0012	0.0011	0	cha	4.9023	0.0006	0.0001	0.0004	0.0003	0
dunt	5.1968	0.0018	0.0005	0.0011	0.0010	0	mieus	4.8405	0.0004	0.0001	0.0003	0.0002	0
od	5.1781	0.0033	0.0009	0.0019	0.0017	0	ochis	4.7118	0.0002	0.0000	0.0002	0.0001	0
si	5.1214	0.0186	0.0136	0.0030	0.0037	0	no	4.6579	0.0005	0.0002	0.0004	0.0003	0
mun	5.0508	0.0018	0.0005	0.0012	0.0010	0	lieu	4.6264	0.0002	0.0001	0.0002	0.0001	0
funt	5.0045	0.0008	0.0002	0.0006	0.0005	0	uausist	4.6239	0.0002	0.0000	0.0001	0.0001	0
reis	4.9249	0.0046	0.0012	0.0033	0.0026	0	espiel	4.6180	0.0004	0.0001	0.0003	0.0002	0
seignurs	4.9082	0.0009	0.0002	0.0006	0.0005	0	laissa	4.6063	0.0001	0.0000	0.0001	0.0001	0
rei	4.8912	0.0038	0.0010	0.0027	0.0022	0	dolans	4.5675	0.0003	0.0001	0.0002	0.0002	0
a	-4.8186	0.0246	0.0328	0.0050	0.0065	0	chi	4.5667	0.0009	0.0003	0.0006	0.0005	0
droit	-4.8320	0.0001	0.0009	0.0002	0.0006	0	toute	4.5588	0.0009	0.0004	0.0002	0.0004	0
qui	-4.8793	0.0037	0.0101	0.0032	0.0050	0	cief	4.4868	0.0007	0.0002	0.0005	0.0004	0
mon	-4.9032	0.0003	0.0023	0.0006	0.0015	0	ainc	4.4662	0.0008	0.0002	0.0005	0.0004	0
et	-4.9212	0.0093	0.0352	0.0195	0.0201	0	mais	4.4656	0.0052	0.0023	0.0014	0.0023	0
sont	-4.9557	0.0003	0.0028	0.0009	0.0019	0	ceual	4.4543	0.0006	0.0002	0.0005	0.0003	0

Table 1: Scriptometric profiles for the Anglo-Norman (left) and Picard groups (right, without the Northern Lorraine subgroup), giving the 25 most characteristic forms (in positive or negative), rounded to 4 decimals

insight as to how clusters were constituted. To do so, the `catdes` function of the `FactoMineR` package by Francois Husson will be used.

The profiles for Anglo-Norman (table 1) shows known features of this *scripta*, like "the replacement of Standard Medieval French (SMF) *o* or *ou* in all positions by *u*", "the retention of *ei* where *SMF* develops *oi*", and "the retention of dentals in 12th-century texts"[23, p. 45-46]. Some are not usually cited: the use of *e* (not *et*), for instance, or *al* (not *au*). The Picard group is also distinctively characterized by its palatalizations, its possessive of 1st and 2nd pers. pl. without -*s* at the singular regime case or nominative plural (*no*, *vo*), the use of *tout*/*tous* (not *tuit*) at the masc. pl. nom., as well as the feminine *toutes*, or the finales in -*s* instead of -*z*.

4 Further research

For the future of this research, an important aspect is the constitution of a corpus more homogeneous in terms of editorial practice. The extension of the corpus, by the addition of new witnesses, would make possible more focused analyses, with, for instance, more restricted chronological limits. The study of the relevance, both from a mathematical and philological point of view, of other metrics, is also a lead for future improvements. It has been shown here, that, though interesting results on the grouping of the witnesses of literary texts can be obtained, their stratified nature remains an obstacle, causing some witnesses to switch groups according to either the presumed *scripta* of their scribe, or the language of the author of the original text. Finding a more satisfying way to account for this phenomenon would be paramount to the scriptometric study of the tradition of medieval literary texts.

References

[1] Bennett, Philip E., 2003, "Le Normand, le picard et les koïnés littéraires de l'épopée aux XII^{e} et XIII^{e} siècles", *Bien Dire et Bien Aprandre*, 21, p. 43-56.

[2] Camps, Jean-Baptiste, 2016, *La 'Chanson d'Otinel': édition complète du corpus manuscrit et prolégomènes à l'édition critique*, dir. Dominique Boutet, thèse de doct., Paris-Sorbonne, DOI: 10.5281/zenodo.1116736.

[3] Dees, Anthonij, Van Reenen, Pieter and De Vries, Johan A, 1980, *Atlas des formes et des constructions des chartes françaises du XIII^{e} siècle*, Tübingen, DOI: 10.1515/9783111328980.

[4] Dees, Anthonij, 1985, "Dialectes et scriptae à l'époque de l'ancien français", *Revue de Linguistique Romane*, 49-193, p. 87–117.

[5] Dees, Anthonij, Dekker, Marcel, Huber, Onno and Van Reenen-Stein, Karin, 1987, *Atlas des formes linguistiques des textes littéraires de l'ancien français*, Tübingen, DOI: 10.1515/9783110935493.

[6] Duval, Frédéric, 2017, "Pour des éditions numériques critiques", *Médiévales*, to be published.

[7] Duval, Frédéric, 2009, *Le français médiéval*, Turnhout.

[8] *Geste: un corpus de chansons de geste*, ed. Jean-Baptiste Camps, 2016-..., Paris, `http://github.com/Jean-Baptiste-Camps/Geste`.

[9] Goebl, Hans, 1995, "Les scriptae françaises III. Normandie", *Les différentes langues romanes et leurs régions d'implantation du Moyen Âge à la Renaissance*, ed. Günter Holtus, *et al.*, Berlin, New York.

[10] Goebl, Hans, 2003, "Regards dialectométriques sur les donées de l''Atlas linguistique de la France' (ALF): Relations quantitatives et structures de profondeur", *Estudis Romànics*, 25, p. 59-120.

[11] Goebl, Hans, 2008, "Sur le changement macrolinguistique survenu entre 1300 et 1900 dans le domaine d'oïl: une étude diachronique d'inspiration dialectométrique", *Dialectologia*, 1.

[12] Goebl, Hans, 2011, "L'aménagement scripturaire du Domaine d'Oïl médiéval à la lumière des calculs de localisation d'Anthonij Dees effectués en 1983: une étude d'inspiration scriptométrique", *Medioevo romanzo*, Seminario 2011: Il problema della scripta, Venezia, `http://www.medioevoromanzo.it/modules/content/index.php?id=14`.

[13] Lebart, Ludovic, Morineau, Alain and Piron, Marie, 1995, *Statistique exploratoire multidimensionnelle*, Paris.

[14] Lusignan, Serge, 2004, *La langue des rois au Moyen Âge: le français en France et en Angleterre*, Paris.

[15] McIntosh, Angus, 1975, "Scribal profiles from Middle English texts", *Neuphilologische Mitteilungen*, 76, p. 218-235.

[16] McIntosh, Angus, 1974, "Towards an inventory of Middle English scribes", *Neuphilologische Mitteilungen*, 75, p. 602-624.

[17] Monfrin, Jacques, 2001, "Le mode de tradition des actes écrits et les études de dialectologie", *Études de philologie romane*, Geneva, p. 145–173.

[18] Morin, Yves-Charles, 2007, "Histoire du corpus d'Amsterdam: le Traitement des données dialectales", *Le Nouveau Corpus d'Amsterdam: actes de l'atelier de Lauterbad, 23-26 février 2006*, ed. Pierre Kunstmann and Achim Stein, Stuttgart, p. 9-27.

[19] Nerbonne, John, and Heeringa Wilbert, 1997, "Measuring dialect distance phonetically", *Workshop on Computational Phonology, Special Interest Group of the* ACL, p. 11–18.

[20] Remacle, Louis, 1948, *Le Problème de l'ancien wallon*, Liège, URL: `http://books.openedition.org/pulg/338`.

[21] Segre, Cesare, 1976, "Critique textuelle, théorie des ensembles et diasystème", *Bulletin de la classe des lettres et des sciences morales et politiques de l'Académie royale de Belgique*, 62, p. 279-92.

[22] Séguy, Jean, 1973, "La dialectométrie dans l'Atlas linguistique de la Gascogne", *Revue de Linguistique romane*, 37, p. 1-24.

[23] Short, Ian, 2007, *Manual of Anglo-Norman*, London.

[24] Strauss, Trudie, and Maltitz, Michael Johan von, 2017, "Generalising Ward's Method for Use with Manhattan Distances", *Plos One*, 12-1, e0168288, DOI: 10.1371/journal.pone.0168288.

[25] Stutzmann, Dominique, 2011, "Paléographie statistique pour décrire, identifier, dater… Normaliser pour coopérer et aller plus loin?", *Kodikologie und Paläographie im digitalen Zeitalter 2*, ed. Franz Fischer, *et al.*, Norderstedt, p. 247-277, `https://halshs.archives-ouvertes.fr/halshs-00596970/`.

[26] Tyssens, Madeleine, 1990, "Typologie de la tradition des textes épiques: les poèmes français", *Memorias de la Real Academia de Buenas Letras de Barcelona*, 22, p. 433-446.

[27] Völker, Harald, 2003, *Skripta und Variation: Untersuchungen zur Negation und zur Substantivflexion in altfranzösischen Urkunden der Grafschaft Luxemburg (1237-1281)*, Tübingen (doct. diss., Univ. of Trier).

A Corpus

Sources: AND = *Anglo-Norman Source Texts*, ed. David A. Trotter, William Rothwell, Geert De Wilde, and Heather Pagan, Aberystwyth and Swansea, 2001, `http://www.anglo-norman.net/sources/`. GESTE [8]. NCA = *Nouveau Corpus d'Amsterdam: corpus informatique de textes littéraires d'ancien français (ca 1150-1350)*, ed. Anthonij Dees, Achim Stein, Pierre Kunstmann, and Martin Dietrich Gleßgen, Stuttgart, `http://www.uni-stuttgart.de/lingrom/stein/corpus`. OTA = *The University of Oxford Text Archive*, ed. University of Oxford IT Services, s. d., `http://ota.ox.ac.uk/`. TFA = *Textes de français ancien*, ed. Pierre Kunstmann and Mark Olsen, 2003, Ottawa, `http://artfl-project.uchicago.edu/content/tfa`. WIKIS = *Wikisource*, ed. Wikimedia Foundation, `http://en.wikisource.org/`.

We follow, when they exist, the identifier given in Möhren, Frankwalt, and Miller, Elena, 2010, *DEAFBiblEl*, Heidelberg, `http://www.deaf-page.de/bibl_neu.php`..

Source	DEAF	ms base	Ed	placeWit	dateWit	placeText	dateText
TFA	AdenBuevH	Ars. 3142	Henry, 1953	Paris	1290pm10	flandr	1275
OTA	AimeriD	BL Roy. 20 B.XIX	Demaison, 1852	bourg	1270ca	nil	1210pm10
NCA+TFA	Aiol1NDeb	BnF fr. 25516	Normand et al., 1877	pic	1275pm25	pic	1160ca
TFA	Aiol2N	BnF fr. 25516	Normand et al., 1877	pic	1275pm25	pic	1210pm10
OTA	AliscW	Ars. 6562	Wienbeck et al., 1903	pic	1213pm13	pic	1190pm10
NCA+TFA	AmAmD	BnF fr. 860	Dembowski, 1969	lorrsept	1275pm25	nil	1200ca
AND	AmAmOctF	BL Roy. 12 C.XII	Fukui, 1990	agn	1335ca	agn	1190pm10
GESTE	Asprem C	Clerm.-Fer. AD 1F2	Camps	agn	1250pm16	agn	1180pm10
GESTE	Asprem P4	BnF, NAF 5094	Albarran & Camps	agn	1200pm20	agn	1180pm10
NCA	AyeB	BnF fr. 2170	Borg, 1967	nil	1300ca	norm	1200ca
TFA	BaudSebC	BnF fr. 12552	Crist, 2002	lorr	1387pm13	pic	1365ca
NCA	CharroiSch A1*	BnF fr. 774	Schoesler	frc	1263pm13	nil	1150pm17
NCA	CharroiSch A2*	BnF fr. 1449	Schoesler	frc	1263pm13	nil	1150pm17
NCA	CharroiSch A3*	BnF fr. 368	Schoesler	lorr	1325pm25	nil	1150pm17
NCA	CharroiSch A4*	Trivulz. 1025	Schoesler	frc	1283pm17	nil	1150pm17
NCA	CharroiSch B1*	BL Royal 20D XI	Schoesler	Paris	1335ca	nil	1150pm17
NCA	CharroiSch B2*	BnF fr. 24369-70	Schoesler	Paris	1335ca	nil	1150pm17
NCA	CharroiSch C*	Boul.-s.-M., BM 192	Schoesler	art	1295	nil	1150pm17
NCA	CharroiSch D*	BnF fr. 1448	Schoesler	lorrmérid	1275pm25	nil	1150pm20
NCA	CharroiSch fr.*	BnF NAF 934	Schoesler	nil	1250pm50	nil	1150pm17
TFA	ChGuillM	BL Add. 38663	McMillan, 1949	agn	1250pm10	agn	1150pm16
TFA	CourLouisLe	BnF fr. 1449	Lepage, 1978	frc	1262pm13	nil	1150pm16
AND	DestrRomeF2	Hann. IV.578	Formisano, 1990	agn	1290pm10	agn	1250pm10
NCA	ElieB*	BnF fr. 25516	P. Bloem	pic	1275pm25	pic	1190pm10
TFA	EnfGarB*	BnF fr. 1460	A. Kostka, 2002	nil	1450pm10	pic	1300ca
GESTE	Fier-V	BAV Reg. lat. 1616	Camps	StBrieuc	1317	nil	1190ca
GESTE	FloovG	Montp., F. Méd. 441	Guessard, 1858	bourg	1325pm25	Sud-Est	1190pm10
NCA	FlorenceW	BnF NAF 4192	Wallenskoeld, 1907	Est	1300ca	pic	1213pm13
NCA	FlorOctOctV	Bodl. Hatton 100	Vollmoeller, 1883	pic	1290pm10	pic	1275pm25
NCA	GirVianeE	BL Roy. 20 B XIX	Van Emden, 1977	bourg	1270ca	champmérid	1210pm10
NCA	GormB	Brux., BR port. II 181	Bayot, 1931	agn	1213pm13	frc	1125pm25
NCA	GuibAndrM	BL Roy. 20 B XIX	Melander, 1922	bourg	1270ca	frc	1210pm10
GESTE	GuiBourgG	Tours, BM 937	Guessard, 1858	nil	1250pm50	nil	1230ca
AND	HornP-C	Cambr. Ff.VI.17	Pope, 1955	agn	1225pm25	agn	1170ca
AND	HornP-O	Bodl. Douce 132	Pope, 1955	agn	1250pm10	agn	1170ca
GESTE	MacaireAl2B	fragm. Loveday	Baker, 1915	agn	1250pm50	nil	1250pm50
TFA	MonGuill1C1	Ars. 6562	Cloetta, 1906	pic	1213pm13	picmérid	1150pm16
TFA	MonGuill1C2	Boul.-s.-M., BM 192	Cloetta, 1906	art	1295	picmérid	1180ca
TFA	MonRaincB	Ars. 6562	Bertin, 1973	pic	1213pm13	pic	1190pm10
WikiS	MortAymC	BL Roy. 20 B.XIX	Couraye, 1884	bourg	1270ca	nil	1213pm13
NCA	OrsonP	BnF NAF 16600	Paris, 1899	lorr	1290pm10	picmérid	1225ca
GESTE	OtinC A	Reg. lat. 1616	Camps	StBrieuc	1317	Nord-Est?	nil
GESTE	OtinC B	Bodmer 168	Camps	agn	1275pm25	Nord-Est?	nil
GESTE	OtinC M	BnF NAF 5094	Camps	agn	1200pm20	Nord-Est?	nil
Divers	PelCharlB	BL Roy. 16 E.VIII	Bonafin, 1987	agn	1290pm10	agn	1175pm25
NCA	PriseCordD	BnF fr. 1448	Densusianu, 1896	Meuse	1262pm13	lorr	1200ca
TFA	PriseOrabR1	BnF fr. 774	Régnier, 1986	Nord-Est	1262pm13	Nord-Est	1190pm10
NCA	RCambr1M	BnF fr. 2493	Meyer et al., 1882	pic	1225pm25	Nord-Est	1190pm10
NCA	RCambr2M	BnF fr. 2493	Meyer et al., 1882	Nord	1275pm25	Nord-Est	1190pm10
NCA	RolS	Bodl. Digby 23	Segre, 1971	agn	1137pm13	Nord-Ouest	1100ca